Computing the Human Interactome

Jing Zhang[1,2,3,†], Ian R. Humphreys[4,5,†], Jimin Pei[1,2,3,†], Jinuk Kim[6], Chulwon Choi[6], Rongqing Yuan[1,2,3], Jesse Durham[1,2,3], Siqi Liu[3,7,8], Hee-Jung Choi[6], Minkyung Baek[6], David Baker[4,5,9], Qian Cong[1,2,3,#]

[1]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[3]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[4] Department of Biochemistry, University of Washington, Seattle, WA, USA.
[5] Institute for Protein Design, University of Washington, Seattle, WA, USA.
[6] Department of Biological Sciences, Seoul National University, Seoul, KR.
[7]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[8] Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA.
[9] Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

[†]Contributed equally

[#]Correspondence: qian.cong@utsouthwestern.edu.

Abstract

Protein-protein interactions (PPI) are essential for biological function. Recent advances in coevolutionary analysis and Deep Learning (DL) based protein structure prediction have enabled comprehensive PPI identification in bacterial and yeast proteomes, but these approaches have limited success to date for the more complex human proteome. Here, we overcome this challenge by 1) enhancing the coevolutionary signals with 7-fold deeper multiple sequence alignments harvested from 30 petabytes of unassembled genomic data, and 2) developing a new DL network trained on augmented datasets of domain-domain interactions from 200 million predicted protein structures. These advancements allow us to systematically screen through 200 million human protein pairs and predict 18,316 PPIs with an expected precision of 90%, among which 5,578 are novel predictions. 3D models of these predicted PPIs nearly triple the number of human PPIs with accurate structural information, providing numerous insights into protein function and mechanisms of human diseases.

Main text

Detecting the interacting partners of proteins and determining the 3D structures of protein complexes are essential to understanding protein function (*1*–*3*). Large-scale experimental methods such as yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (APMS) have been used to identify protein-protein interactions (PPI) across the human proteome (*4*–*7*). While powerful, these methods detect PPIs in non-physiological conditions and are associated with considerable false-positive and false-negative rates (*8*, *9*). This is illustrated by the fact that although the human interactome is estimated to contain 74,000 to 200,000 PPIs (*10*), the experimental data in various PPI databases, UniProt (*11*), BioGrid (*12*), and STRING (*13*), suggest over 1 million (M) human PPIs may exist. Additionally, high-confidence PPIs from these databases show low consistency (**Fig. 1A**), with only 3988 PPIs regarded as confident by all three.

Computational approaches have been developed to complement experimental studies in resolving the human interactome. They predict PPIs based on homology to known interacting partners, quality of the predicted interfaces, and functional associations (*14*–*21*). Coevolutionary analysis (residues at PPI interfaces must coevolve to maintain their interactions) has been combined with 3D structure prediction methods, such as AlphaFold (AF) and RoseTTAFold (RF), to identify PPIs on a proteome-wide scale in bacteria and yeast (*22*, *23*). The probability for any protein pair to interact is estimated by 1) creating multiple sequence alignments (MSAs) of orthologs of the two proteins in many species, 2) concatenating sequences of the same species to build paired MSAs (pMSAs), and 3) determining the probabilities for residues in the first protein to interact with residues in the second based on coevolution and complex structure modeling (*24*). While similar approaches have been used to detect interactions between selected human proteins (*18*, *25*–*27*), a *de novo* proteome-wide PPI screen in humans is challenging due to the daunting computational scale, and the small number of high eukaryotic genomes that limits the statistical power of coevolution between human proteins.

We set out to overcome these challenges and systematically identify PPIs in humans. To enhance the statistical power of coevolutionary analysis, we reasoned that it should be possible to harness the petabytes of untapped genomic sequence data in draft eukaryotic genomes and genomic reads. To analyze hundreds of millions human protein pairs efficiently, we sought to develop a fast but accurate Deep Learning (DL) network for PPI prediction by 1) optimizing the architecture of state-of-the-art DL networks and 2) augmenting the training set of experimentally determined PPIs in the PDB with a much larger distillation set of domain-domain interactions (DDIs) harvested from accurately predicted

monomer structures in the AlphaFold protein structure Database (AFDB) (*28*), based on the assumption that DDI interfaces should resemble PPI interfaces in coevolutionary and physicochemical properties.

Harnessing untapped genomic sequence data

Biologists typically rely on databases with protein sequences annotated from genomes (such as Uniref) or metagenomic assemblies (such as BFD (*29*) and MGnify (*30*)). However, the majority of available eukaryotic genomes, especially for higher Eukaryotes, have not been annotated with protein sequences. Identifying protein-coding regions in eukaryotic genomes is non-trivial due to the complicated and frequent lineage-specific rules of translation initiation and mRNA splicing (*31*). Hence, it is not a routine practice to annotate and deposit the protein-coding sequences for draft eukaryotic genomes: among the 36,840 eukaryotic genomes available at NCBI in June, 2024, only 7,355 (20%) were associated with annotated proteins. Moreover, assembling diploid, heterozygous, and repeat-rich eukaryotic genomes is challenging without long-read sequencing (*32*). The NCBI Sequence Read Archive (SRA) database (*33*) has over 30 petabytes of shotgun sequencing reads, which have not been assembled into contigs or analyzed to predict the proteins they encode.

To enrich the coevolutionary information in our MSAs, we mined the NCBI genome and SRA databases to extract whole genome and whole transcriptome datasets for 21,414 diverse Eukaryotes, focusing on Chordates and Arthropods, two large phylums of higher Eukaryotes. We selected one representative genomic dataset per species, including 1) 2,424 species whose proteome sequences were previously annotated from genomes, 2) 6,863 species with draft genomes but no protein annotations, and 3) 12,128 species with whole genome or whole transcriptome shotgun reads. We developed a bioinformatic pipeline (see **supplemental Methods M4**) to assemble protein-coding sequences from each type of dataset, utilizing splicing-site-aware sequence aligners (*34*, *35*) and reference protein sets from model organisms (**Fig. 1B**). All predicted protein sequences were aligned to their human orthologs, if available, and we used reciprocal best-hit criteria (*36*) to distinguish orthologs from paralogs. We name the resulting alignments omicMSAs, because they are directly derived from genomic data. Our omicMSAs include sequences from 21,414 species spanning 9,905 genera, 2,727 families, and 626 orders, significantly expanding the taxonomic diversity of available protein sequences in UniRef100, which only contains 3,082 species (**Fig. 1C**).

We reasoned that the rich evolutionary information in omicMSAs should improve the ability of DL networks to distinguish true PPIs from false ones. We tested this hypothesis using benchmarks derived from PPI databases. There are 75,739, 77,147, and 68,761 confident human PPIs from UniRef, BioGrid (multi-validated), and STRING (physical interactions with combined score > 700), respectively. However, only 3,988 (5.3%–5.8%) PPIs are shared by all (**Fig. 1A**), from which we selected 100 non-overlapping positive control sets, each containing 36 PPIs. The negative control set contains 36,000 random pairs of human proteins showing no evidence for their interactions, and its much larger size than the positive control sets allows performance evaluation at a low signal-to-noise ratio (1:1000) that approximates the situation of proteome-wide PPI screen (74k–200k true PPIs out of 200M pairs). We compared omicMSAs against MSAs built by other commonly used strategies. The first is HHblits (*37*) against UniRef, a widely used approach to prepare MSA inputs for RF2 and AF2. HHblits by default filters the output MSAs at 90% sequence identity, and we named this strategy UniRef90. Filtering MSAs before building pMSAs significantly reduces the number of taxa that can be paired; to increase pMSA depth, we disabled this filter and termed the alternative strategy UniRef100.

Metagenomic sequences are frequently used to obtain deeper MSAs and improve structure prediction (*38*). We used the ColabFold MSA pipeline (*39*) to combine UniRef and metagenomic sequences.

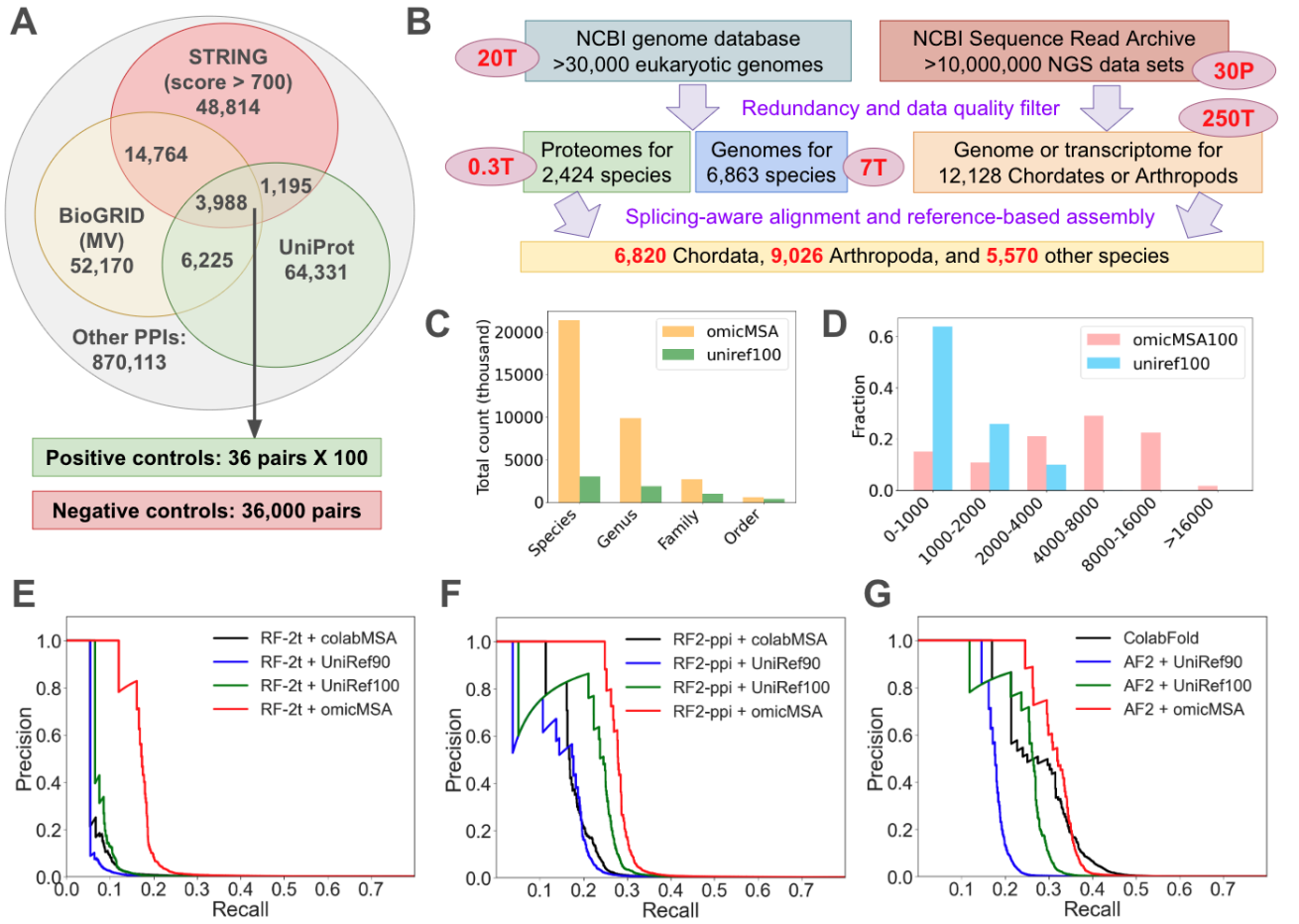

Figure 1. OmicMSA increases PPI identification accuracy. **(A)** Intersections of confident PPIs from different databases and our strategy to select the benchmark sets. **(B)** Strategies to assemble protein sequences from genomic data at NCBI. The boxes show statistics of different datasets, and the ovals by the boxes indicate their volume (T: terabytes, P: petabytes). Texts in purple by the arrows describe the data processing procedures. **(C)** Taxonomic diversity of our MSAs derived from genomic data and eukaryotic protein sequences in UniRef100. **(D)** Distribution of pMSA depth (filtered at 100% identity) for protein pairs in the benchmark sets. **(E-G)** Precision and recall curves (averaged over the 100 positive control sets) of different methods in distinguishing the positive from the negative controls (signal:noise = **1:1000**). OmicMSA consistently improves the performance of different methods, such as **(E)** RF-2t (*23*), **(F)** RF2-ppi, and **(G)** AF2.

We found that omicMSA led to the best performance for all the tested methods (**supplemental fig S24**), including RF-2t used in yeast PPI screen (*23*), the newly developed RF2-ppi in this study, and AF2, in distinguishing true PPIs from random pairs (red curves in **Fig. 1E-G**). The ColabFold MSA (black curves in **Fig. 1E-G**) improves over UniRef100 when used with AF2, but it did not perform better with other tools. Metagenomic data, mostly sequenced from environmental samples, do not contain sequences of higher Eukaryotes and are not annotated with the assumption of mRNA splicing (*40*); thus, they do not necessarily improve human PPI prediction. In contrast, we focused on assembling sequences from genomic data of higher Eukaryotes with pipelines aware of mRNA splicing. All the tools we tested were trained using MSAs built from UniRef and metagenomic sequences, and they are not adapted to work with omicMSA. Thus, the superior performance of omicMSA demonstrates the power of stronger evolutionary signals gleaned from a much wider range

of taxa. The available eukaryotic genomic data is constantly growing (~20% more species per year), providing even greater potential in the future.

RoseTTAFold2-PPI: a DL network for rapid PPI identification

Because the physicochemical properties of residue-residue contacts at PPI interfaces are expected to differ from contacts within tightly packed protein monomers (*2*, *3*), AlphaFold-multimer (AFmm) (*41*) was trained with PPIs in the PDB to improve its ability to model 3D structures of protein complexes. However, such training may not improve an DL network's ability on the distinct task of distinguishing true PPIs from false ones — instead, this might bias the network to predict PPIs between random pairs without strong coevolutionary signals. Indeed, AFmm does not perform well in distinguishing true PPIs from random pairs at a signal-to-noise ratio of 1:1000 (**Fig. 2D**). Trained on PPIs from the PDB and optimized for speed, we developed RF2-lite (*22*) to distinguish true PPIs from random pairs. RF2-lite is 20 times faster than AF2, but shows considerably lower accuracy than AF2 (**Fig. 2D**). We reason that one factor limiting the accuracy of current methods is the amount of training data. At 30% sequence identity, there are 24,358 clusters of PPIs (hetero-oligomers only) in the PDB (December, 2023), among which many pairs are human proteins or their homologs (MMseqs e-value < 0.00001); after removing these complexes to avoid information leakage for human PPI prediction, only 13,231 clusters remain (**Fig. 2B left**).

We hypothesized that extracted DDIs from the 200M AF2 models in the AFDB (*42*) could significantly enlarge the training dataset for PPI predictors and improve their performance. DDIs within the same protein should resemble PPIs (*43*, *44*). Domains are structural and evolutionary units that are recombined to create new proteins throughout evolution (*45*), and the interfaces between domains closely resemble those between distinct protein partners (*43*, *44*) . We previously developed a method to segment AF2 models into domains based on structural features (inter-residue distances and predicted aligned errors (PAEs)) and homology to previously classified domains from PDB entries (*46*, *47*). We repurposed this method to integrate structural features and domain annotations of UniProt entries from InterPro (**Fig. 2A** and **supplemental Methods M2.2**). We found 12.4M multi-domain proteins among the 53.7M AFDB models filtered at 50% sequence identity (*48*). These models contain 22.6M high-quality (mean PAE within domains < 8) domain pairs from the same proteins. We focused on pairs with at least 25 inter-residue contacts (distance < 6 Å) and high confidence in their interaction (mean inter-domain PAE < 8), resulting in 1M DDIs. Clustering these DDIs at 30% sequence identity resulted in 237,919 clusters (**Fig. 2B right**), 10-fold larger than the PPI training set.

We used both the PDB PPI and AFDB DDI sets to develop RF2-ppi (**Fig. 2C**) based on the RF2 architecture (*49*), which integrates the MSA, inter-residue, and 3D structure features through attention-based blocks. Additionally, we included monomeric 3D structures to ensure that RF2-ppi learns principles of protein folding; we added negative controls – random protein or domain pairs from the same organism with no evidence of being functionally related – to ensure that RF2-ppi learns to distinguish interacting partners from random pairs. The monomers, positive DDIs, negative DDIs, positive PPIs, and negative PPIs were combined at a 2:1:1:1:1 ratio to generate the training dataset. We explored a number of strategies to optimize the performance of RF2-ppi (see **supplemental Methods M3.2**); we found that adding the DDI training set and removing the 3D structure features and losses significantly improved its performance (**supplemental figs. S10 and S11**). The lightweight RF2-ppi performed better than RF2 for PPI identification (**supplemental fig. S15**), which again suggests that optimizing predicted 3D structures of protein complexes does not necessarily help to distinguish true PPIs from random pairs. This may be because weak human PPIs are frequently

mediated by simple motifs (single helices or intrinsically disordered regions (IDRs)), while our training sets are dominated by interactions with larger interfaces. Direct integration of 3D features may drive the DL network to emphasize on the geometric and chemical complementarity at the PPI interfaces instead of the coevolutionary signals essential for predicting weak PPIs. The 3D structure module of AF2 is not directly integrated into the Evoformer, possibly allowing AF2 to properly balance coevolutionary signals and physicochemical properties.

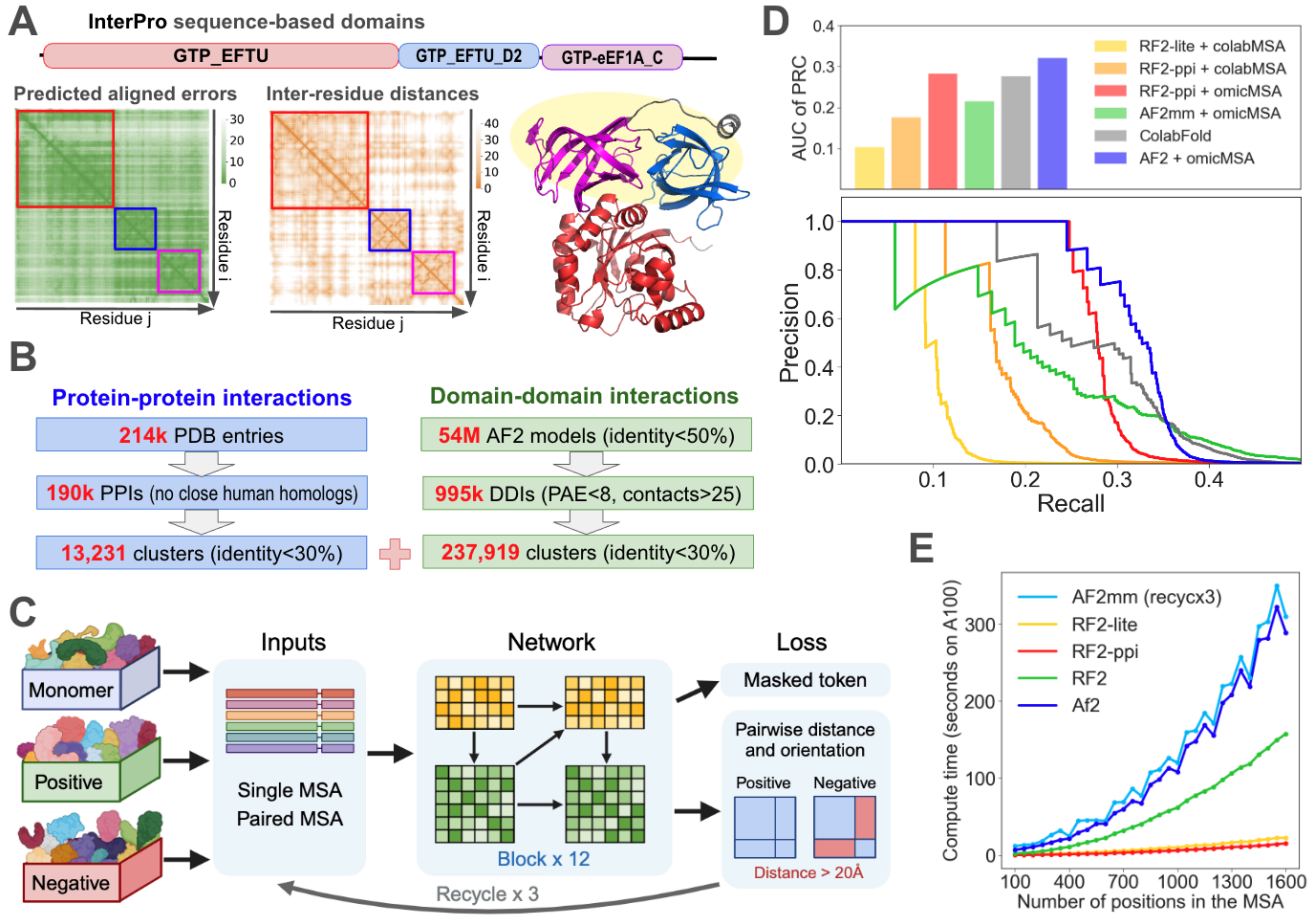

Figure 2. A DL network for PPI prediction, trained using datasets augmented with predicted DDIs. (A) Illustration of our domain segmentation protocol. Top: domains annotated by InterPro; left: PAE matrix; middle: inter-residue distance matrix; right: 3D structure. In both matrices, the parsed domains are marked by boxes colored the same as in the 3D structure. The two domains in yellow shade belong to the DDI training set; their interaction with the 3rd domain did not pass our inter-domain PAE cutoff. **(B)** Statistics of the PPI and DDI training datasets. **(C)** Architecture and training routine for RF2-ppi. **(D)** Precision and recall curves (below) and the area under the curves (above) for different methods in their ability to distinguish true PPIs from false ones (signal: noise = 1:1000). **(E)** Compute time as a function of input protein size for different methods.

We compared the performance of RF2-ppi with the ColabFold pipeline (skyblue in **Fig. 2D**) widely used by the scientific community (*39*) and other methods. Using area under the precision and recall curve (AUCPR) as a metric, RF2-ppi (orange in **Fig. 2D**) showed a 1.7-fold improvement over our previous method, RF2-lite (gold in **Fig. 2D**). When combined with omicMSAs, RF2-ppi (red in **Fig. 2D**) slightly outperforms the ColabFold pipeline in its ability to identify true PPIs and is 20 times faster than AF2/ColabFold (**Fig. 2E**). AFmm (green in **Fig. 2D**), despite its superior performance in modeling 3D structures of protein complexes, shows significantly worse performance than RF2-ppi in distinguishing true PPIs from random pairs. The inferior performance of AFmm is related to the low signal-to-noise

ratio (1:1000) associated with *de novo* PPI screens; at a signal-to-noise ratio of 1:10, AFmm will outperform the other methods. The best performance is achieved when AF2 is used in combination with omicMSAs (deepblue in **Fig. 2D**). We used both RF2-ppi and AF2 to identify and model human PPIs: the former allows us to carry out our screen on a proteome-wide scale, while the latter is needed to obtain high-quality 3D structure models of confident PPIs.

Proteome-wide identification of human PPIs

Equipped with omicMSA and RF2-ppi, we set out to comprehensively predict the human interactome. We took AF2 models for the human proteome (20,504 proteins) from the AFDB and identified domains based on structural compactness (*46*) and evolutionary conservation (*50*). We included these domains and relatively conserved residues in the inter-domain linkers (conservation > 25% quantile of residues in domains) in our screen; excluding regions that are poorly conserved and lack rigid 3D structures improves the performance of our PPI screen pipeline (**supplemental fig. S23**). To fit into limited GPU memory, we split large proteins into segments with few inter-residue contacts and flexible relative orientation (see **supplemental Methods M5.1**) between them. In total, we screened 191M pairs comprised of 19,528 proteins, excluding a small fraction (4.8%) of proteins due to protein size limits or the lack of compact structures and conserved motifs.

We began with an unbiased systematic search for PPIs across the 191M protein pairs. To make the search more tractable, we focused on pairs in the same cellular compartment (CC) based on keywords annotated by UniProt. We analyzed 53.8M pairs sharing CC annotations and 57.4M pairs involving proteins without CC annotations sequentially through Direct Coupling Analysis (DCA) (*50*, *51*), RF2-ppi, and AF2 (**Fig. 3A**). Although DCA, a statistical method to detect coevolution, shows significantly worse performance than the state-of-the-art DL tools, it is 5-fold faster than RF2-ppi and we used it to select 43.6M (40%) pairs. Based on our benchmark, most (over 80%) pairs showing high interaction probability by AF2 will pass this DCA pre-filter (**supplemental table S4**). This *de novo* pipeline (see **Fig. 3B** for performance evaluation and cutoff at each step) predicted 6,966 PPIs with an expected precision of 90%. We complemented these results by carrying out a second screen incorporating the extensive information about human PPIs derived primarily from high throughput experiments. Evaluating such noisy experimental data with our *in silico* pipeline allows us to identify a smaller set of PPIs at high confidence. Prior evidence of interactions allowed us to use lower RF2-ppi and AF2 cutoffs while maintaining the same level of expected precision. We predicted 10,627 PPIs among the 5.23M genetically interacting pairs from STRING (**Fig. 3A**) and 12,832 PPIs from the 1.15M pairs with evidence for physical interactions from UniProt, BioGRID, and STRING. The *de novo* (unbiased) and evidence-guided pipelines allowed us to predict 18,307 PPIs at 90% precision and 30.8% recall (see **supplemental Methods M5.6**).

We compared our predictions against PPIs from other databases (**Fig. 3C**). We reasoned that the fraction of PPIs from another source that can be identified by our *in silico* screen ($f_{AI/DB}$) reflects the precision of PPIs in that database, which can be computed as $pre_{DB} = f_{AI/DB} / rec_{AI}$, $rec_{AI} = 0.308$. PPIs with orthologous PDB templates are mostly (93%) true. However, not all contacting chains in PDB entries interact in physiological conditions; some instead result from experimental conditions and crystal packing (*52*, *53*). The estimated precision of other PPI databases is low (4% – 12%), but the confident subset selected by each database has higher precision (20% – 30%, **Fig. 3D**). Compared to PPIs in databases, predicted PPIs display a much stronger tendency to participate in the same biological processes and locate to the same subcellular components (**Fig. 3E**). Thus, our predictions

can be used to infer the functions and subcellular localities of poorly characterized proteins by finding their well-characterized partners.

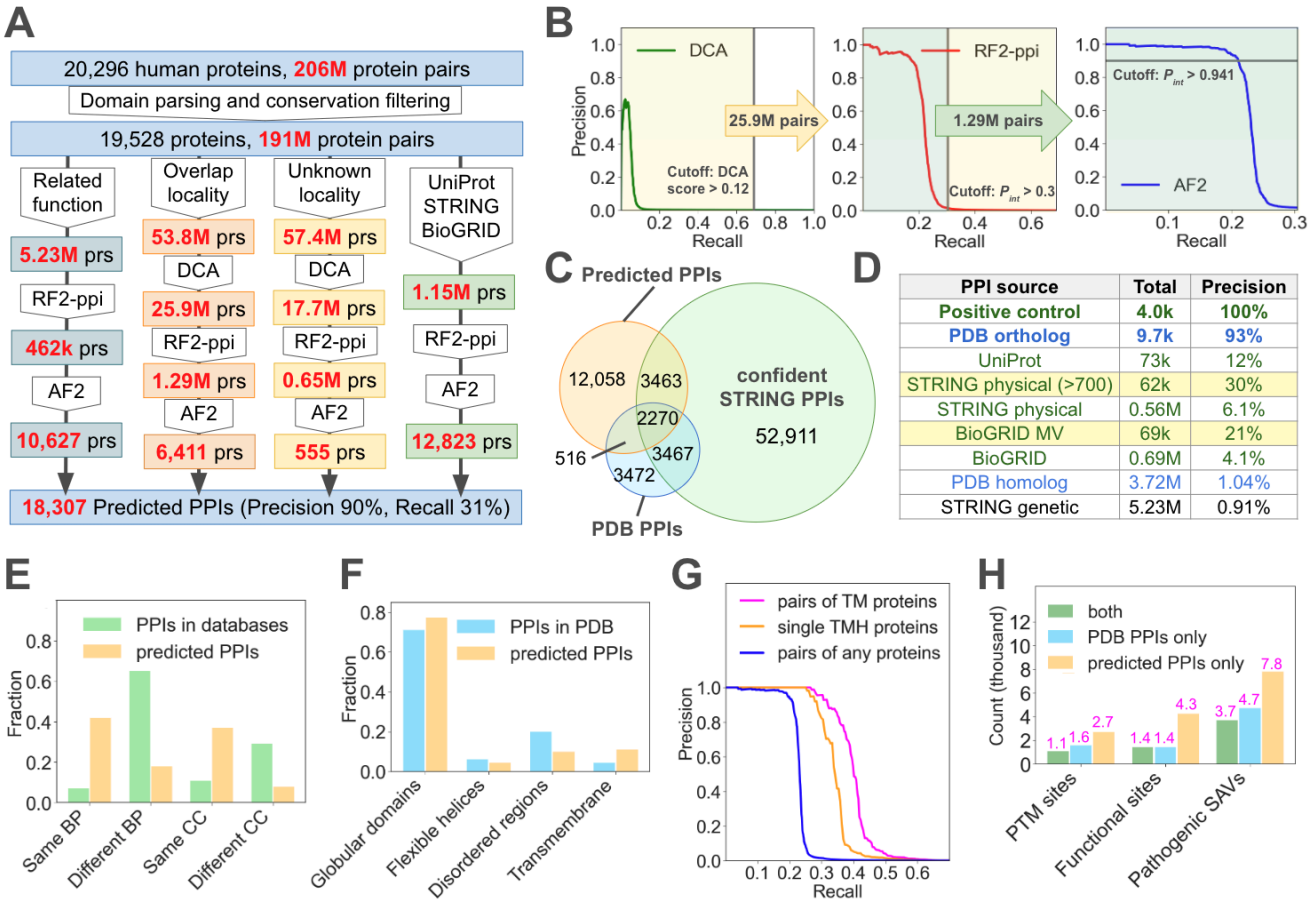

Figure 3. Human Proteome-wide PPI identification. (A) PPI screen strategies. Key methods are shown in white connectors and numbers of selected pairs after key steps are shown in colored squares. **(B)** Performance of our *de novo* pipeline applied to colocalized protein pairs. The grey line indicates the recall (vertical) or precision (horizontal) cutoff (corresponding score labeled by the line) at each stage; number of pairs selected at each stage is shown in the arrow. **(C)** A Venn diagram showing the overlap between our predictions and PPIs from other sources; such overlap are used to estimate **(D)** the precision of PPIs in other databases: PDB ortholog, with orthologous PDB templates; PDB homolog, with homologous PDB templates; yellow highlights: the confident subsets from STRING and BioGRID. **(E)** Predicted PPIs show a much higher chance to participate in the same biological processes (BP) and localize in the same cellular components (CC) than human PPIs in databases. Same: complete overlap of UniProt keywords; Different: no overlap; partial overlap not shown. **(F)** Fraction of different interface types in predicted and experimentally determined PPIs. Globular domains: >50% interface residues for both; flexible helices: >50% interface residues for one; disordered regions: >50% interface residues for one; transmembrane: >25% interface residues for one. **(G)** Performance of our *de novo* pipeline for TMPs. **(H)** Numbers (labeled on top) of functional sites at PPI interfaces.

We computed 3D structures of predicted PPIs using AF2 and AFmm and selected the best models (see **supplemental Methods M6.2**). We compared the properties of predicted PPIs against experimentally determined PPIs, i.e. 9,725 human PPIs with orthologous templates in the PDB. Similar to PDB complexes, most (76%) predicted PPIs are primarily mediated by globular domains. However, interfaces mediated by flexible helices or IDRs are much rarer among predicted PPIs than in PDB complexes (**Fig. 3F**), consistent with the fact that PPIs with smaller interfaces are harder to predict (**supplemental fig. S28**). A higher fraction of predicted PPI interfaces involves

transmembrane helices (TMHs) than experimentally determined. We wondered if the hydrophobic nature of TMHs make them prone to false positives because DL networks might have learned to pack hydrophobic TMHs against each other even if they do not coevolve or interact *in vivo*. We tested the performance of our pipeline on pairs of transmembrane proteins (TMPs) and found that it predicts a larger fraction (magenta curve versus blue curve in **Fig. 3G**) of their interactions at a high precision, including those mediated by single-TMH proteins. These observations suggest that our pipeline is suitable for identifying interactions between TMPs, which are hard to detect experimentally (*54–56*). The 3D models of predicted PPIs in this study nearly triple the number of human PPIs with high-quality 3D structures. Thousands of post-translational modification (PTM) sites, functional (active and substrate/cofactor binding) sites, and disease-associated single amino acid variations (SAVs) (*57–59*) map to the predicted PPI interfaces (**Fig. 3H**), and thus in-depth analysis of our models will provide numerous mechanistic insights of protein functions.

Biological insights from novel PPIs

Among the 18,307 predictions, 2,786 (15%) PPIs or their orthologs have experimental 3D structures; another 9,953 (54.3%) are reported in PPI databases (UniProt, BioGRID, and STRING physical); the remaining 5,568 pairs (30.4%) are novel compared to these established sources of PPIs (**Fig. 4A**). Many of the novel predictions are indirectly supported by other evidence such as genetic interactions in STRING (3,443 pairs) and homologous (not orthologous) templates in the PDB (2,673 pairs); some have been experimentally tested based on literature search but are not yet propagated into PPI databases. Based on a recent catalog of poorly characterized proteins (*60*), 1,582 predicted PPIs involve proteins of unknown functions: linking them to well-characterized proteins provides shortcuts to investigating their functions. For our in-depth analysis, we prioritized 2,895 PPIs that are absent from PPI databases and lack homologous PDB templates. We classified predicted PPIs into functional categories according to UniProt keywords, and found categories enriched (P-value < 0.01) in such novel PPIs (**Fig. 4B**). Focusing on several such categories, below we illustrate how our results can be used to study protein function. Additionally, many novel predictions involve proteins associated with genetic disorders or cancers (**Fig. 4C-4H**) and potentially offer insights into human disease (see **supplemental Results R2.1 and R2.6)**.

G protein-coupled receptors (GPCRs)

GPCRs are a large family of TMPs that detect extracellular signals and mediate cellular responses (*61*). They play vital roles in human physiology and are important drug targets (*62*). We detect novel interactions between GPCRs and their putative downstream signaling molecules and ligands or modulators. For example, we predict an interaction between GPR143 and HPS1 (**Fig. 4I**). GPR143 is an atypical GPCR localized in endolysosomes and melanosomes, while HPS1 is a component of a guanine nucleotide exchange factor (GEF) complex that activates Ras-related GTPase RAB32 and RAB38 (*63*). The GPR143-HPS1 interaction is intriguing because both proteins are expressed in melanosomes, involved in melanosome biogenesis, and associated with albinism (*64–67*). In addition, we detect a potential ligand or modulator for ADGRF5 (**Fig. 4J**), an adhesion GPCR critical for lung surfactant homeostasis (*68*); its predicted interaction with SFTA2 via its extracellular domain suggests that this secreted small protein, predominantly expressed in lung (*69*), can be the ligand or modulator for ADGRF5. Further, we predict an interaction between C5AR2 (a GPCR) and ELANE, a serine protease (**Fig. 4K**). The former is known to be a receptor for C5a anaphylatoxin peptide and plays a role in chemotaxis and inflammation (*70*), while the latter inhibits C5a-dependent chemotaxis and

neutrophil function (*71*). This predicted C5AR2-ELANE interaction and their opposite roles in chemotaxis suggests that ELANE is a negative modulator of C5AR2.

We also observe examples of predicted interactions between pairs of GPCRs as well as between GPCRs and other TMPs that may function together to deliver complex cellular responses. For example, we predict an interaction between GPR35 and GPRC5C (**Fig. 4L**). GPR35 is involved in chemotaxis of macrophage and inflammation responses (*72*–*74*), while GPRC5C promotes the dormancy of hematopoietic stem cells (*75*). This predicted heterodimer, with a non-conventional dimer interface (between TMH1&2 of GPR35 and TMH6&7 of GPRC5C), may unravel novel signaling pathways in blood cell differentiation. In addition, HTR6, a GPCR mediating neurotransmission (*76*), is predicted to interact with CDH22, a single-TMH protein in the cadherin family involved in cell adhesion in the brain (*77*, *78*) (**Fig. 4M**). Such an interaction provides a potential link between GPCR signaling and cell adhesion. Finally, we predict an interaction between an orphan GPCR, GPR152 (*79*), and TMEM42 (**Fig. 4N**), an uncharacterized TMP showing homology to a variety of transporters according to HHpred (*80*). The GPR152-TMEM42 interaction points to the possibility of transporter activity regulation by a GPCR.

Immunity proteins

The evolutionary arms race between hosts and pathogens drive hosts to utilize a wide range of mechanisms to defend against infectious bacteria and viruses (*81*). We predict 265 novel PPIs involving immunity proteins, many of which mediate innate immunity signaling pathways. For instance, we predict an intriguing interaction between a deubiquitinase, OTUD4, and an E3 ubiquitin ligase, ASB8 (**Fig. 4O**). OTUD4 removes ubiquitin chains from MAVS, an important innate immune adapter to detect viral RNA, preventing its degradation (*82*); in contrast, ASB8 introduces polyubiquitins to MAVS's downstream protein kinase TBK1, promoting its degradation (*83*). The interaction of these two proteins with opposing roles might help maintain a balance in the regulation of protein turnover during innate immune responses. We predict interactions between FCN1, an extracellular receptor for pathogen recognition (*84*), and several extracellular enzymes (**Fig. 4P**), including two tryptases (TPSAB1 and TPSAB2) and a hyaluronidase (HYAL1). These interactions suggest a potential role of FCN1 in innate immune responses in the extracellular matrix, opening up new directions to investigate its functions. Interestingly, we predict an interaction between an interferon-induced TMP (IFITM1) and a SNARE protein VAMP8 (**Fig. 4Q**). IFITM1 inhibits the entry of many viruses, including SARS-CoV, into the host cells (*85*), while VAMP8 is involved in the fusion of autophagosomes and lysosomes (*86*, *87*). This suggests a potential mechanism for IFITM1's antiviral activity by recruiting VAMP8 and promoting the delivery of virus cargo in the autophagosomes to lysosomes for degradation.

Other novel PPI predictions involve proteins with roles in the differentiation and activation of specialized immune cells. For instance, PLPP6 is an enzyme contributing to neutrophil activation via dephosphorylation of presqualene diphosphate, a potent inhibitor of this process (*88*, *89*). We predict an interaction between PLPP6 and an unknown protein LOC122513141 (**Fig. 4R**). Both proteins are located in the Endoplasmic Reticulum (ER) membrane, and this predicted interaction suggests that LOC122513141 could serve as a regulator of PLPP6 function. In addition, KLRG1 is a TM receptor that inhibits the activity of natural killer cells and effector T cells (*90*, *91*). The extracellular domain of KLRG1 is predicted to interact with TLR3, a key player in recognizing double-stranded RNA and inducing immune responses (*92*) (**Fig. 4S**). Such an interaction suggests a potential mechanism for KLRG1's function as an immunity inhibitor. Finally, we predict an interaction between two remote homologs, TYROBP and CD3D (**Fig. 4T**). The former mediates the activation of a variety of immune cells (*93*–*95*), while the latter transmits signals from T-cell receptors and activates T-cells (*96*). The

TYROBP-CD3D interaction suggests TYROBP' role in T-cell activation, potentially an example of the complicated cooperation between cell surface receptors in activating various immune cells.

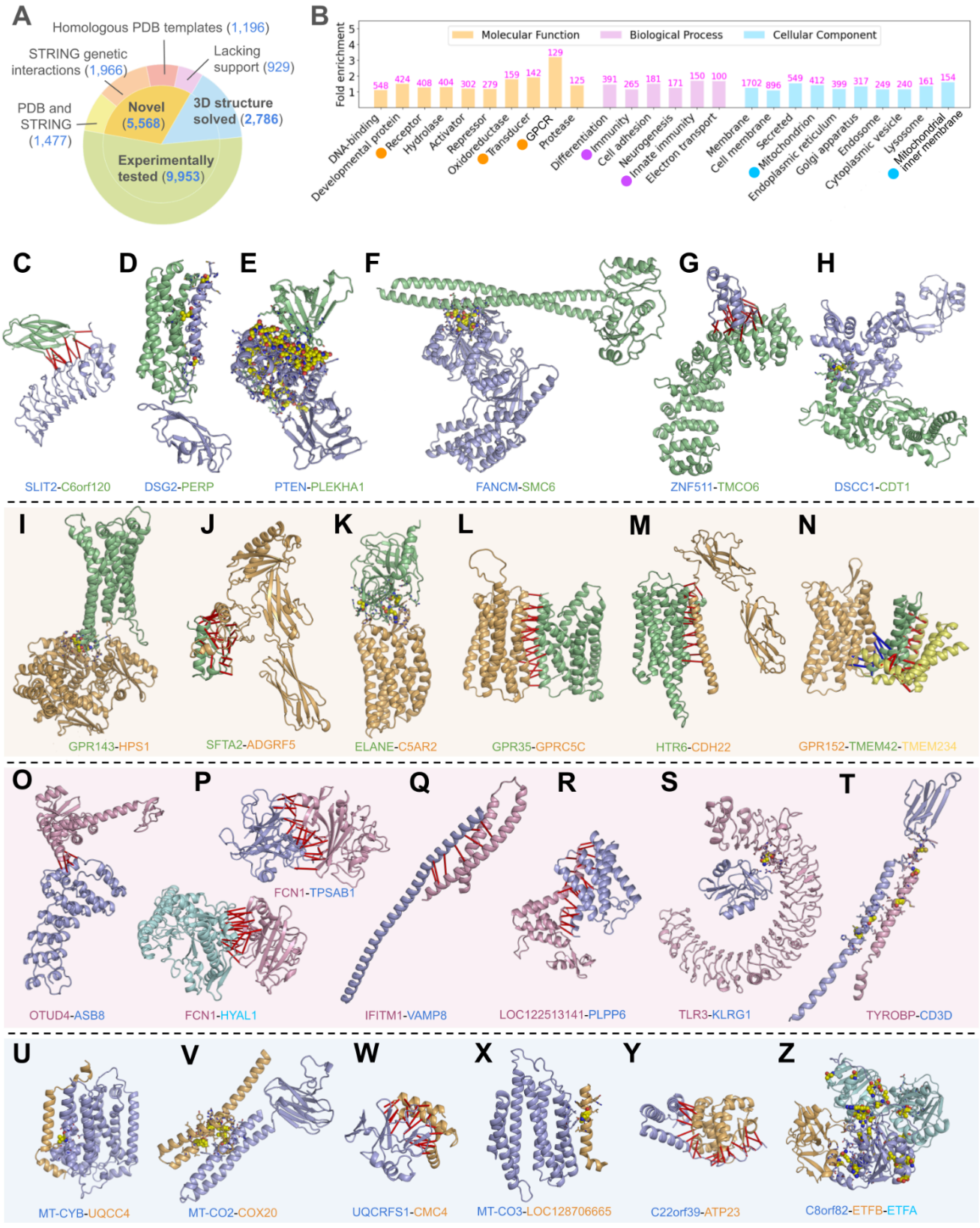

Figure 4. Examples of novel functional insights. (A) A pie chart showing the fraction of predicted PPIs with experimental evidence. 3D structure solved: with orthologous PDB template; experimentally tested: present in UniProt, STRING (physical), or BioGRID databases, but 3D structures not resolved; the remaining novel PPIs are further partitioned by whether they show evidence of genetic interactions, have homologous templates, or both. **(B)** Functional categories (UniProt keywords) that are enriched (P-value < 0.01) with novel PPIs (numbers labeled above the bars). Dots by the labels mark categories discussed in the main text. **(C-H)** PPIs involving cancer-related proteins. **(I-N)** PPIs involving GPCRs. **(O-T)** PPIs related to immunity. **(U-Z)** Mitochondrial PPIs. **(C-Z)** Yellow spheres indicate interface residues associated with disease-causing SAVs; otherwise, red and blue bars connect residues with their most confidently predicted inter-protein contacts.

Mitochondrial targeting proteins

More than half of PPIs not found in databases (1,697 out of the 2,895) involve TMPs. These TMPs are frequently located in cellular organelles, such as mitochondrion, ER, Golgi apparatus, endosome, and lysosome (**Fig. 4B**). Many PPIs we predict in the mitochondria are related to the assembly of mitochondrial complexes that constitute the respiratory electron transport chain (ETC). For example, we detect interactions between the complex III component MT-CYB (*97*) and the complex III assembly factor UQCC4 (*98*) (**Fig. 4U**), and between the complex IV subunit MT-CO2 (*99*) and the complex IV assembly factor COX20 (*100*) (**Fig. 4V**). Our predictions also reveal several proteins of unknown functions that may serve as additional assembly factors of the ETC complexes. For example, we link CMC4 (**Fig. 4W**), LOC128706665 (**Fig. 4X**), C22orf39 (**Fig. 4Y**), and C8orf82 (**Fig. 4Z**) to components or assembly factors of complex III, IV, V, and the electron transfer flavoprotein complex, respectively. Among them, LOC128706665 was annotated as “alternative protein MKKS” in UniProt because its coding region overlaps with the 5’ UTR of the molecular chaperone MKKS that primarily resides in the centrosome (*101*). However, our prediction suggests that LOC128706665 should be a novel mitochondrial protein not related to MKKS.

New multi-protein complexes

By integrating our predictions and PPIs derived from orthologous PDB templates, we identify 379 predicted multi-protein complexes, wherein each component interacts with at least two other components (**supplemental Methods M6.3**). We classify PPIs in these complexes into three categories: (1) with experimental structures, (2) confident entries of PPI databases; (3) no or weak experimental support. We focus on putative complexes dominated by PPIs lacking confident experimental support (3rd category). These complexes (**Fig. 5**) are frequently scaffolded by proteins with long IDRs or flexible helices that interact with multiple components. They likely adopt flexible 3D structures and are hard for experimental characterization; predicting their existence and modeling their 3D structures open new directions for future studies.

Tubulin polyglutamylase (TPG) is an enzyme complex responsible for adding polyglutamate chains to the glutamate residues in the C-terminal tail of tubulin, a modification important for the regulation of microtubule functions (*102*, *103*). In addition to the catalytic subunit TTLL1, the TPG complex is known to contain four other subunits: TPGS1, TPGS2, LRRC49 and NICN1 (*104*). We predict 3 additional subunits of TPG: TBC1D19, CSTPP1, and SANBR, each interacting with multiple known TPG components (**Fig. 5A**). TBC1D19 is the central component of the complex, interacting with 5 other subunits (TTLL1, LRRC49, NICN1, TPGS1, and SANBR). TBC1D19 and SANBR directly interact with the catalytic subunit TTLL1 via interfaces apart from its ATP binding and catalytic sites (*105*), possibly affecting the activity of TTLL1 by allosteric regulation.

Another predicted complex consisting of NOL10, UTP25, DDX47, ESF1, ABT1, and DDX49, could link the regulation of basal transcription with ribosome biogenesis (**Fig. 5B**). The two processes should be coordinated to ensure that transcribed mRNA can be efficiently translated into proteins. The central component, NOL10, is a poorly characterized nucleolar protein made of an N-terminal beta-propeller domain and C-terminal flexible helices. ABT1 and its interacting partner ESF1 are suggested to regulate basal transcription (*106*, *107*), while UTP25 and DDX47 are implicated in ribosome biogenesis (*108*). The connection between this complex and the ribosome is also supported by its predicted interactions with ribosomal proteins via ESF1, DDX47, and NOL10; the N-domain of NOL10 is even part of an experimental ribosome structure (PDB: 7MQ8) (*109*). Future studies of this putative complex may reveal new insights about the coordination between the two fundamental processes, transcription and translation.

We predict several complexes made of proteins associated with the biogenesis of cilia and flagella, two cellular organelles sharing the same molecular machineries during their formation (*110*). The flagellum enables a sperm to swim towards and fertilize the egg (*111*), linking these complexes to sperm motility and reproduction. The first predicted complex includes CATIP, RIIAD1, MDH1B, CFAP91, AKAP14, AK9, and MORN5, among which CFAP91, an intrinsically disordered protein, serves as a scaffold to assemble all but one (RIIAD1) other proteins (**Fig. 5C**). Involvement of this complex in flagella biogenesis and sperm motility is supported by several facts: 1) CFAP91 and CATIP have been shown to participate in cell skeleton organization during cilia/flagella biogenesis (*112*–*115*); 2) AK9 affects sperm mobility (*116*); and 3) defects in CATIP and AK9 are associated with asthenozoospermia (*115*, *116*). We predict two additional complexes which involve proteins implicated in sperm biogenesis. One consists of CFAP69, LRGUK, and SPEF2 (**Fig. 5D**); the other is made of CFAP65, CFAP70, MYCBPAP, and ARMC3 (**Fig. 5E**). These complexes are likely parts of a larger machinery, connected via an interaction between CFAP65 and LRGUK.

New components of known protein complexes

We identify 83 potential new components of known human protein complexes (**supplemental Methods M6.3**) in the complex portal database (*117*). Each new component is predicted to interact with at least two subunits of a known complex and primarily interact (>50% of predicted PPIs) with this complex. For example, LYAR, a nuclear protein involved in rRNA processing (*118*), is predicted to interact with two telomere maintenance complex (TMC) proteins, DKC1 and GAR1 (**Fig. 5F**). Additional components of TMC include NOP10, NHP2, and WRAP53, and they participate in the processing and trafficking of the RNA template (*119*–*123*) used to synthesize telomeric DNA (*124*). Multiple components of the TMC, including DKC1, GAR1, NOP10, and NHP2, have been shown to participate in rRNA processing and ribosome biosynthesis (*119*, *125*–*127*). The binding of LYAR to these components might channel the TMC to process rRNA instead of telomeric RNA template.

We predict a new component to the Glycosylphosphatidylinositol-N-acetylglucosaminyltransferase (GPI-GnT) complex, which catalyzes the first step of GPI anchor biosynthesis (*128*). The GPI-GnT complex consists of seven subunits: PIGA, PIGC, PIGH, PIGP, PIGQ, PIGY, and DPM2 (*128*), and its 3D structure remains unresolved. ARV1, a TMP found to be related to GPI biosynthesis (*129*–*131*), is predicted to interact with PIGC and PIGQ. Our 3D model of the GPI-GnT complex with ARV1 (**Fig. 5G**) shows that a hydrophobic helix (residues 422-442) in the catalytic subunit, PIGA, is positioned in parallel to the membrane and docked onto a heterotrimer of the 2-pass TMPs: PIGP, PIGY, and DPM2. ARV1 (5 TMHs) is positioned in between PIGC (8 TMHs) and PIGQ, and several hydrophobic

helices of PIGQ lie in parallel to the membrane. These subunits form a ring-like structure with a large cavity in the middle of the membrane that could facilitate substrate binding or transport of the product.

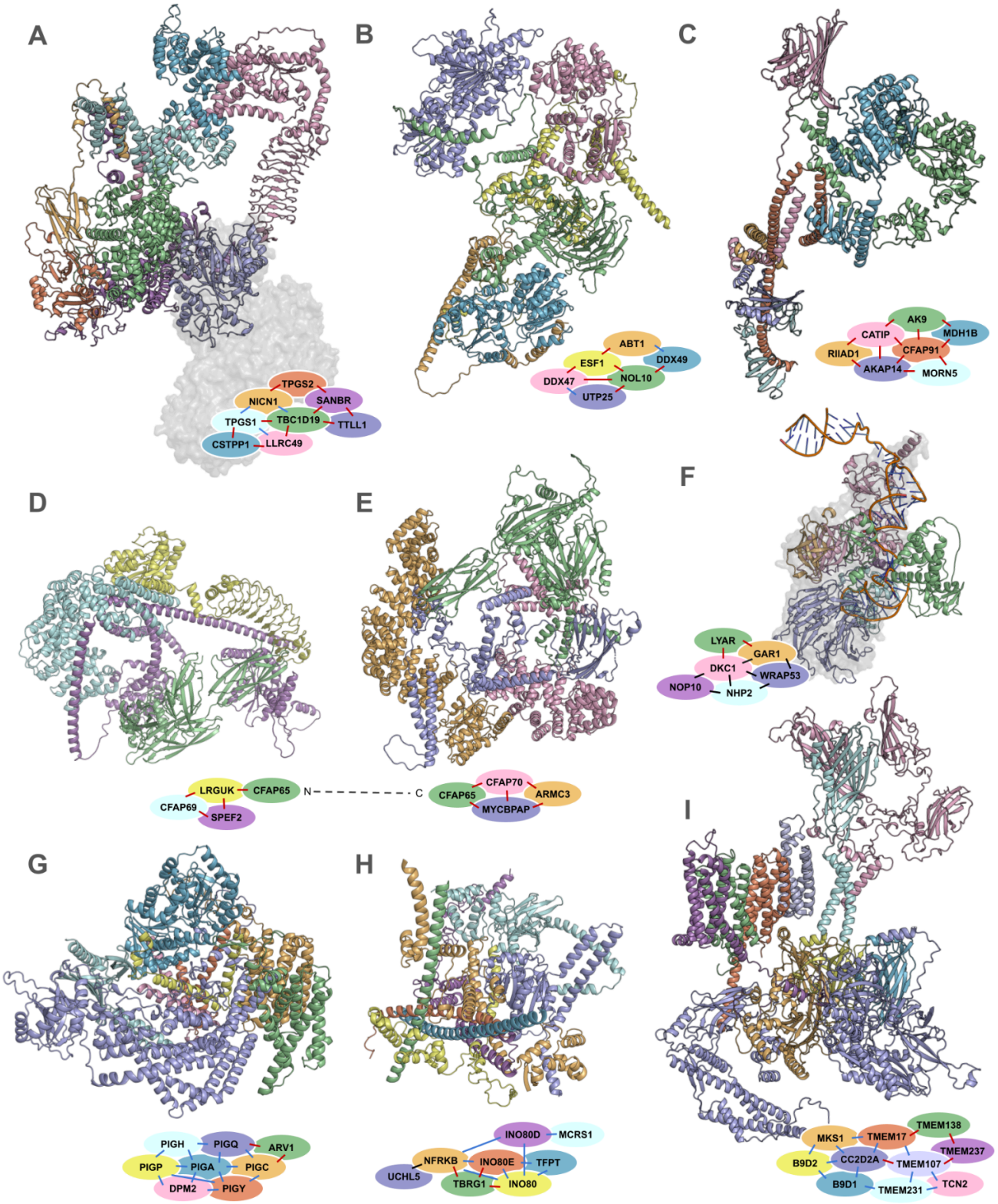

Figure 5. Examples of novel protein complexes (A-E) and predicted new components to known complexes (F-I). Each panel contains the 3D structure of a complex and a network graph. Nodes in the graph

are labeled by gene names and colored the same as the corresponding subunits in the 3D structure. Edges in the graph represent PPIs and are colored as follows: black, PPIs with experimental structures; blue, confident PPIs from databases; red: novel PPIs or with weak experimental support. **(A)** The tubulin polyglutamylase complex. An experimental structure (PDB: 5jh7) of the catalytic subunit and its substrate (tubulins) is shown as surface representation and colored in gray. **(B)** A complex connecting basal transcription regulation with ribosome biosynthesis. **(C-E)** Complexes related to cilia and flagella biogenesis. **(F)** Adding LYAR to the telomere maintenance complex; an experimental structure of this complex (PDB: 8ouf) is shown as surface representation and colored in gray. **(G)** Adding ARV1 to a GPI anchor biosynthesis complex. **(H)** Adding TBRG1 to the INO80 chromatin remodeling complex. **(I)** Adding TMEM138 to the transition zone complex.

We predict multiple interactions between components of the INO80 chromatin remodeling complex (CRC) (*132*), which allowed us to model a subcomplex comprised of the regulatory subunits (INO80D, INO80E, NFRKB, TFPT, MCRS1, and UCHL5). This model reveals that the regulatory subcomplex interacts with the N-terminal IDR of the CRC core subunits, INO80, while the C-terminal domains of INO80 interact with other core CRC subunits to perform the chromatin remodeling function (*133*). The regulatory subunits likely help to coordinate the complicated roles of CRC, including transcription regulation, DNA replication, and DNA repair (*134*, *135*). Recently, it was suggested that TBRG1 could be a new subunit of the CRC based on its interaction with INO80 (*136*). Consistent with this study, we predict that TBRG1 interacts with multiple CRC subunits through a long N-terminal helix, not only INO80, but also TFPT, NFRKB, and INO80E **(Fig. 5H)**. TBRG1, as well as TFPT, are related to the regulation of cell cycle and cell growth (*137*), potentially linking the CRC function to these processes.

Finally, we predict a new component of the MKS transition zone complex (TZC) located between the basal body and axoneme of cilia and flagella (*138*, *139*). Predicted PPIs between TZC components enabled us to model a large complex **(Fig. 5I)** spanning the extracellular (TCN2 and TMEM231), membrane (TMEM17, TMEM107, TMEM138, and TMEM237), and cytoplasmic space (MKS1, B9D1, B9D2, and CC2D2A). TMEM138 has been suggested to localize to the transition zone based on immunofluorescence assays and is associated with the same ciliopathies as other transition zone proteins (*140*). We further predict TMEM138 as a subunit of TZC, sandwiched between TMEM17 and TMEM237 in our computed model. We included one copy of each TZC subunit in our model; however, it is likely that these proteins will homo-oligomerize to form a much larger complex which constitutes the transition zone.

Conclusions

Despite significant advances in DL methods for modeling the 3D structures of protein complexes, distinguishing the 74k-200k true PPIs amongst the 200M pairs of human proteins has remained a grand challenge. Here, we tackle this challenge by leveraging the largest sequence and structure datasets available and focusing on two key innovations: 1) strengthening the coevolutionary signals between interacting proteins using 7-fold deeper MSAs directly assembled from genomic data and 2) developing a new and fast DL network, trained on 10 times larger datasets derived from predicted DDIs, to differentiate true PPIs from random protein pairs. These improvements enabled us to systematically identify human PPIs on a proteome-wide scale, predicting over 18k interactions with an expected precision of 90%. Notably, these predictions include more than 5.5k novel PPIs, particularly among TMPs that are challenging to characterize experimentally. These predictions will offer numerous valuable insights into human biology and diseases, as demonstrated by our biological vignettes. The power of our approach will continue to grow as more sequence and structure data become available and DL techniques advance. Integrating the rapidly improving computational

approaches with experimental studies, we show that the nearly complete characterization of the human interactome is within reach.

Availability

We present our predictions and intermediate results at http://prodata.swmed.edu/humanPPI, which allows users to easily navigate through our findings and accelerate future discoveries. We share our omicMSAs, training datasets, trained weights of RF2-ppi, predicted interaction probabilities, and 3D models of protein complexes at: https://conglab.swmed.edu/humanPPI/humanPPI_download.html. We deposited the RF2-ppi network at GitHub: https://github.com/CongLabCode/RoseTTAFold2-PPI. To seamlessly use our omicMSA to build protein complexes with AF2 or AFmm, please use our Colab notebook at: https://colab.research.google.com/drive/1suhoIB5q6xn0APFHJE8c1eMiCuv9gCk_

Acknowledgment

The authors thank Nick V. Grishin, Jian Zhou, Rohith Krishna, Frank Dimaio, Edin Muratspahic, and Helen H. Hobbs for inspiring discussions and insightful suggestions. We also thank Luki Goldschmidt and Aaron Guillory for computing resource management, and Linda Stewart and Lance Stuart for logistical support. QC is a Southwestern Medical Foundation Endowed Scholar and DB is a Howard Hughes Medical Institute Investigator. This research is supported by I-2095-20220331 and V-I-0004-20230731 from the Welch Foundation, 1K99AI180984-01A1, 5-R01-HL-159946-03, and contract No. 75N93022C00036 from NIH, Department of Health and Human Services, Bill & Melinda Gates Foundation Investments INV-010680 and INV-043758, Novo Nordisk A/S contract RDC 2022-002902, NRF/MSIT (RS-2023-00210147, RS-2024-00440824, and RS-2024-00407331), IITP/MSIT (RS-2023-00220628), and Perlmutter grant NERSC award BER-ERCAP0022018 for access to the Perlmutter high-performance computing resources.

Reference

1. V. S. Rao, K. Srinivas, G. N. Sujini, G. N. S. Kumar, Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics* **2014**, 147648 (2014).
2. J. Zhang, J. Durham, Qian Cong, Revolutionizing protein-protein interaction prediction with deep learning. *Curr. Opin. Struct. Biol.* **85**, 102775 (2024).
3. J. Durham, J. Zhang, I. R. Humphreys, J. Pei, Q. Cong, Recent advances in predicting and modeling protein-protein interactions. *Trends Biochem. Sci.* **48**, 527–538 (2023).
4. T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Wanamaker, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, M. Vidal, A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
5. J. H. Morris, G. M. Knudsen, E. Verschueren, J. R. Johnson, P. Cimermancic, A. L. Greninger, A. R. Pico, Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.* **9**, 2539–2554 (2014).
6. K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, D. Choi, A. G. Coté, M. Daley, S. Deimling, A. Desbuleux, A. Dricot, M. Gebbia, M. F. Hardy, N. Kishore, J. J. Knapp, I. A. Kovács, I. Lemmens, M. W. Mee, J. C. Mellor, C. Pollis, C. Pons, A. D. Richardson, S. Schlabach, B. Teeking, A. Yadav, M. Babor, D. Balcha, O. Basha, C. Bowman-Colin, S.-F. Chin, S. G. Choi, C. Colabella, G. Coppin, C. D'Amata, D. De Ridder, S. De Rouck, M. Duran-Frigola, H. Ennajdaoui, F. Goebels, L. Goehring, A. Gopal, G. Haddad, E. Hatchi, M. Helmy, Y. Jacob, Y. Kassa, S. Landini, R. Li, N. van Lieshout, A. MacWilliams, D. Markey, J. N. Paulson, S. Rangarajan, J. Rasla, A. Rayhan, T. Rolland, A. San-Miguel, Y. Shen, D. Sheykhkarimli, G. M. Sheynkman, E. Simonovsky, M. Taşan, A. Tejeda, V. Tropepe, J.-C. Twizere, Y. Wang, R. J. Weatheritt, J. Weile, Y. Xia, X. Yang, E. Yeger-Lotem, Q. Zhong, P. Aloy, G. D. Bader, J. De Las Rivas, S. Gaudet, T. Hao, J. Rak, J. Tavernier, D. E. Hill, M. Vidal, F. P. Roth, M. A. Calderwood, A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
7. K. Yugandhar, S. Gupta, H. Yu, Inferring Protein-Protein Interaction Networks From Mass Spectrometry-Based Proteomic Approaches: A Mini-Review. *Comput. Struct. Biotechnol. J.* **17**, 805–811 (2019).
8. J. P. Mackay, M. Sunde, J. A. Lowry, M. Crossley, J. M. Matthews, Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007).
9. C. M. Deane, Ł. Salwiński, I. Xenarios, D. Eisenberg, Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349–356 (2002).
10. K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, M. Vidal, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
11. UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
12. R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, M. Tyers, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).
13. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2023:

protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

14. D. Petrey, H. Zhao, S. J. Trudeau, D. Murray, B. Honig, PrePPI: A Structure Informed Proteome-wide Database of Protein-Protein Interactions. *J. Mol. Biol.* **435**, 168052 (2023).
15. G. Lasso, S. V. Mayer, E. R. Winkelmann, T. Chu, O. Elliot, J. A. Patino-Galindo, K. Park, R. Rabadan, B. Honig, S. D. Shapira, A Structure-Informed Atlas of Human-Virus Interactions. *Cell* **178**, 1526–1541.e16 (2019).
16. M. I. Parker, J. E. Meyer, E. A. Golemis, R. L. Dunbrack, Delineating the RAS Conformational Landscape. *Cancer Res.* **82**, 2485–2498 (2022).
17. Q. Xu, R. L. Dunbrack Jr, ProtCID: a data resource for structural information on protein interactions. *Nat. Commun.* **11**, 711 (2020).
18. D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, A. Elofsson, Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023).
19. H. Zhao, D. Petrey, D. Murray, B. Honig, ZEPPI: Proteome-scale sequence-based evaluation of protein-protein interaction models. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2400260121 (2024).
20. B. de Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agaugué, G. Meiffren, F. Pradezynski, B. F. Faria, T. Chantier, M. Le Breton, J. Pellet, N. Davoust, P. E. Mangeot, A. Chaboud, F. Penin, Y. Jacob, P. O. Vidalain, M. Vidal, P. André, C. Rabourdin-Combe, V. Lotteau, Hepatitis C virus infection protein network. *Mol. Syst. Biol.* **4**, 230 (2008).
21. M. Gao, D. Nakajima An, J. M. Parks, J. Skolnick, AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).
22. I. R. Humphreys, J. Zhang, M. Baek, Y. Wang, A. Krishnakumar, J. Pei, I. Anishchenko, C. A. Tower, B. A. Jackson, T. Warrier, D. T. Hung, S. B. Peterson, J. D. Mougous, Q. Cong, D. Baker, Essential and virulence-related protein interactions of pathogens revealed through deep learning. *bioRxiv*, doi: 10.1101/2024.04.12.589144 (2024).
23. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, D. Baker, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
24. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
25. J. Zhang, J. Pei, J. Durham, T. Bos, Q. Cong, Computed cancer interactome explains the effects of somatic mutations in cancers. *Protein Sci.* **31**, e4479 (2022).
26. J. Pei, J. Zhang, Q. Cong, Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling. *Bioinformatics* **38**, 4301–4311 (2022).
27. J. Pei, J. Zhang, Q. Cong, Computational analysis of protein-protein interactions of cancer drivers in renal cell carcinoma. *FEBS Open Bio* **14**, 112–126 (2024).
28. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
29. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
30. L. Richardson, B. Allen, G. Baldi, M. Beracochea, M. L. Bileschi, T. Burdett, J. Burgin, J.

Caballero-Pérez, G. Cochrane, L. J. Colwell, T. Curtis, A. Escobar-Zepeda, T. A. Gurbich, V. Kale, A. Korobeynikov, S. Raj, A. B. Rogers, E. Sakharova, S. Sanchez, D. J. Wilkinson, R. D. Finn, MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

31. R. Guigó, Genome annotation: From human genetics to biodiversity genomics. *Cell Genom* **3**, 100375 (2023).
32. W. De Coster, M. H. Weissensteiner, F. J. Sedlazeck, Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
33. K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, C. O'Sullivan, The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
34. H. Li, Protein-to-genome alignment with miniprot. *Bioinformatics* **39** (2023).
35. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
36. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003).
37. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
38. S. Lee, G. Kim, E. L. Karin, M. Mirdita, S. Park, R. Chikhi, A. Babaian, A. Kryshtafovych, M. Steinegger, Petabase-Scale Homology Search for Structure Prediction. *Cold Spring Harb. Perspect. Biol.* **16** (2024).
39. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
40. H. Alexander, S. K. Hu, A. I. Krinos, M. Pachiadaki, B. J. Tully, C. J. Neely, T. Reiter, Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *MBio* **14**, e0167623 (2023).
41. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, *bioRxiv* (2022)p. 2021.10.04.463034.
42. M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
43. S. Jones, A. Marin, J. M. Thornton, Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77–82 (2000).
44. R. Verma, S. B. Pandit, Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions. *PLoS One* **14**, e0220336 (2019).
45. N. Sen, M. S. Madhusudhan, A structural database of chain–chain and domain–domain interfaces of proteins. *Protein Sci.* **31**, e4406 (2022).
46. J. Zhang, R. D. Schaeffer, J. Durham, Q. Cong, N. V. Grishin, DPAM: A domain parser for AlphaFold models. *Protein Sci.* **32**, e4548 (2023).
47. H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, N. V. Grishin, ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).
48. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
49. M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, F. DiMaio, Efficient and accurate prediction of protein structure using RoseTTAFold2, *bioRxiv* (2023)p. 2023.05.24.542179.
50. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
51. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 67–72 (2009).

52. J. Luo, Z. Liu, Y. Guo, M. Li, A structural dissection of large protein-protein crystal packing contacts. *Sci. Rep.* **5**, 14214 (2015).
53. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
54. S. Cao, S. M. Peterson, S. Müller, M. Reichelt, C. McRoberts Amador, N. Martinez-Martin, A membrane protein display platform for receptor interactome discovery. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).
55. A. L. Richards, M. Eckhardt, N. J. Krogan, Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Mol. Syst. Biol.* **17**, e8792 (2021).
56. G. Khazen, A. Gyulkhandanian, T. Issa, R. C. Maroun, Getting to know each other: PPIMem, a novel approach for predicting transmembrane protein-protein complexes. *Comput. Struct. Biotechnol. J.* **19**, 5184–5197 (2021).
57. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
58. H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. S. D. Dietrich, P. P. Fiziev, L. F. K. Kuderna, L. Sundaram, Y. Wu, A. Adhikari, Y. Field, C. Chen, S. Batzoglou, F. Aguet, G. Lemire, R. Reimers, D. Balick, M. C. Janiak, M. Kuhlwilm, J. D. Orkin, S. Manu, A. Valenzuela, J. Bergman, M. Rousselle, F. E. Silva, L. Agueda, J. Blanc, M. Gut, D. de Vries, I. Goodhead, R. Alan Harris, M. Raveendran, A. Jensen, I. S. Chuma, J. E. Horvath, C. Hvilsom, D. Juan, P. Frandsen, F. R. de Melo, F. Bertuol, H. Byrne, I. Sampaio, I. Farias, J. V. do Amaral, M. Messias, M. N. F. da Silva, M. Trivedi, R. Rossi, T. Hrbek, N. Andriaholinirina, C. J. Rabarivola, A. Zaramody, C. J. Jolly, J. Phillips-Conroy, G. Wilkerson, C. Abee, J. H. Simmons, E. Fernandez-Duque, S. Kanthaswamy, F. Shiferaw, D. Wu, L. Zhou, Y. Shao, G. Zhang, J. D. Keyyu, S. Knauf, M. D. Le, E. Lizano, S. Merker, A. Navarro, T. Bataillon, T. Nadler, C. C. Khor, J. Lee, P. Tan, W. K. Lim, A. C. Kitchener, D. Zinner, I. Gut, A. Melin, K. Guschanski, M. H. Schierup, R. M. D. Beck, G. Umapathy, C. Roos, J. P. Boubli, M. Lek, S. Sunyaev, A. O'Donnell-Luria, H. L. Rehm, J. Xu, J. Rogers, T. Marques-Bonet, K. K.-H. Farh, The landscape of tolerated genetic variation in humans and primates. *Science*, doi: 10.1126/science.abn8197 (2023).
59. J. Jänes, M. Müller, S. Selvaraj, D. Manoel, J. Stephenson, C. Gonçalves, A. Lafita, B. Polacco, K. Obernier, K. Alasoo, M. C. Lemos, N. Krogan, M. Martin, L. R. Saraiva, D. Burke, P. Beltrao, Predicted mechanistic impacts of human protein missense variants, *bioRxiv* (2024)p. 2024.05.29.596373.
60. J. J. Rocha, S. A. Jayaram, T. J. Stevens, N. Muschalik, R. D. Shah, S. Emran, C. Robles, M. Freeman, S. Munro, Functional unknomics: Systematic screening of conserved genes of unknown function. *PLoS Biol.* **21**, e3002222 (2023).
61. M. Zhang, T. Chen, X. Lu, X. Lan, Z. Chen, S. Lu, G protein-coupled receptors (GPCRs): advances in structures, mechanisms, and drug discovery. *Signal Transduct Target Ther* **9**, 88 (2024).
62. A. S. Hauser, M. M. Attwood, M. Rask-Andersen, H. B. Schiöth, D. E. Gloriam, Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).
63. A. Gerondopoulos, L. Langemeyer, J.-R. Liang, A. Linford, F. A. Barr, BLOC-3 mutated in Hermansky-Pudlak syndrome is a Rab32/38 guanine nucleotide exchange factor. *Curr. Biol.* **22**, 2135–2139 (2012).
64. B. Bueschbell, P. Manga, A. C. Schiedel, The Many Faces of G Protein-Coupled Receptor 143, an Atypical Intracellular Receptor. *Front Mol Biosci* **9**, 873777 (2022).
65. R. Bakker, E. L. Wagstaff, C. C. Kruijt, E. Emri, C. D. M. van Karnebeek, M. B. Hoffmann, B. P. Brooks, C. J. F. Boon, L. Montoliu, M. M. van Genderen, A. A. Bergen, The retinal pigmentation pathway in human albinism: Not so black and white. *Prog. Retin. Eye Res.* **91**, 101091 (2022).
66. L. Le, J. Sirés-Campos, G. Raposo, C. Delevoye, M. S. Marks, Melanosome Biogenesis in the Pigmentation of Mammalian Skin. *Integr. Comp. Biol.* **61**, 1517–1545 (2021).
67. M. V. Schiaffino, Signaling pathways in melanosome biogenesis and pathology. *Int. J. Biochem. Cell Biol.* **42**, 1094–1104 (2010).
68. F. Kubo, D. M. Ariestanti, S. Oki, T. Fukuzawa, R. Demizu, T. Sato, R. M. Sabirin, S. Hirose, N.

Nakamura, Loss of the adhesion G-protein coupled receptor ADGRF5 in mice induces airway inflammation and the expression of CCL2 in lung endothelial cells. *Respir. Res.* **20**, 11 (2019).

69. R. A. Mittal, M. Hammel, J. Schwarz, K. M. Heschl, N. Bretschneider, A. W. Flemmer, S. Herber-Jonat, M. Königshoff, O. Eickelberg, A. Holzinger, SFTA2--a novel secretory peptide highly expressed in the lung--is modulated by lipopolysaccharide but not hyperoxia. *PLoS One* **7**, e40011 (2012).
70. T. Zhang, K.-Y. Wu, N. Ma, L.-L. Wei, M. Garstka, W. Zhou, K. Li, The C5a/C5aR2 axis promotes renal inflammation and tissue damage. *JCI Insight* **5** (2020).
71. T. Tralau, U. Meyer-Hoffert, J.-M. Schröder, O. Wiedow, Human leukocyte elastase and cathepsin G are specific inhibitors of C5a-dependent neutrophil enzyme release and chemotaxis. *Exp. Dermatol.* **13**, 316–325 (2004).
72. A. Boleij, P. Fathi, W. Dalton, B. Park, X. Wu, D. Huso, J. Allen, S. Besharati, R. A. Anders, F. Housseau, A. E. Mackenzie, L. Jenkins, G. Milligan, S. Wu, C. L. Sears, G-protein coupled receptor 35 (GPR35) regulates the colonic epithelial cell response to enterotoxigenic Bacteroides fragilis. *Commun Biol* **4**, 585 (2021).
73. S. Oka, R. Ota, M. Shima, A. Yamashita, T. Sugiura, GPR35 is a novel lysophosphatidic acid receptor. *Biochem. Biophys. Res. Commun.* **395**, 232–237 (2010).
74. M. De Giovanni, H. Tam, C. Valet, Y. Xu, M. R. Looney, J. G. Cyster, GPR35 promotes neutrophil recruitment in response to serotonin metabolite 5-HIAA. *Cell* **185**, 815–830.e19 (2022).
75. Y. W. Zhang, J. Mess, N. Aizarani, P. Mishra, C. Johnson, M. C. Romero-Mulero, J. Rettkowski, K. Schönberger, N. Obier, K. Jäcklein, N. M. Woessner, M.-E. Lalioti, T. Velasco-Hernandez, K. Sikora, R. Wäsch, B. Lehnertz, G. Sauvageau, T. Manke, P. Menendez, S. G. Walter, S. Minguet, E. Laurenti, S. Günther, D. Grün, N. Cabezas-Wallscheid, Hyaluronic acid-GPRC5C signalling promotes dormancy in haematopoietic stem cells. *Nat. Cell Biol.* **24**, 1038–1048 (2022).
76. D. Marazziti, S. Baroni, F. Borsini, M. Picchetti, E. Vatteroni, V. Falaschi, M. Catena-Dell'Osso, Serotonin receptors of type 6 (5-HT6): from neuroscience to clinical pharmacology. *Curr. Med. Chem.* **20**, 371–377 (2013).
77. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
78. M. Mayer, K. Bercsényi, K. Géczi, G. Szabó, Z. Lele, Expression of two type II cadherins, Cdh12 and Cdh22 in the developing and adult mouse brain. *Gene Expr. Patterns* **10**, 351–360 (2010).
79. S. Lu, W. Jang, A. Inoue, N. A. Lambert, Constitutive G protein coupling profiles of understudied orphan GPCRs. *PLoS One* **16**, e0247743 (2021).
80. J. Söding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–8 (2005).
81. M. Sironi, R. Cagliani, D. Forni, M. Clerici, Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236 (2015).
82. T. Liuyu, K. Yu, L. Ye, Z. Zhang, M. Zhang, Y. Ren, Z. Cai, Q. Zhu, D. Lin, B. Zhong, Induction of OTUD4 by viral infection promotes antiviral responses through deubiquitinating and stabilizing MAVS. *Cell Res.* **29**, 67–79 (2019).
83. Y. Guo, R. Li, Z. Tan, J. Shi, Y. Fu, Y. Song, M. Zhu, L. Zhang, J. Huang, E3 ubiquitin ligase ASB8 negatively regulates interferon via regulating TBK1/IKKi homeostasis. *Mol. Immunol.* **121**, 195–203 (2020).
84. S. Bidula, D. W. Sexton, S. Schelenz, Ficolins and the Recognition of Pathogenic Microorganisms: An Overview of the Innate Immune Response and Contribution of Single Nucleotide Polymorphisms. *J Immunol Res* **2019**, 3205072 (2019).
85. G. Shi, A. D. Kenney, E. Kudryashova, A. Zani, L. Zhang, K. K. Lai, L. Hall-Stoodley, R. T. Robinson, D. S. Kudryashov, A. A. Compton, J. S. Yount, Opposing activities of IFITM proteins in SARS-CoV-2 infection. *EMBO J.* **40**, e106501 (2021).
86. Q. Chen, M. Hao, L. Wang, L. Li, Y. Chen, X. Shao, Z. Tian, R. A. Pfuetzner, Q. Zhong, A. T. Brunger,

J.-L. Guan, J. Diao, Prefused lysosomes cluster on autophagosomes regulated by VAMP8. *Cell Death Dis.* **12**, 939 (2021).

87. H. Huang, Q. Ouyang, M. Zhu, H. Yu, K. Mei, R. Liu, mTOR-mediated phosphorylation of VAMP8 and SCFD1 regulates autophagosome maturation. *Nat. Commun.* **12**, 6622 (2021).
88. K. Fukunaga, M. Arita, M. Takahashi, A. J. Morris, M. Pfeffer, B. D. Levy, Identification and functional characterization of a presqualene diphosphate phosphatase. *J. Biol. Chem.* **281**, 9490–9497 (2006).
89. T. Carlo, H. Kalwa, B. D. Levy, 15-Epi-lipoxin A4 inhibits human neutrophil superoxide anion generation by regulating polyisoprenyl diphosphate phosphatase 1. *FASEB J.* **27**, 2733–2741 (2013).
90. B. Müller-Durovic, A. Lanna, L. P. Covre, R. S. Mills, S. M. Henson, A. N. Akbar, Killer Cell Lectin-like Receptor G1 Inhibits NK Cell Function through Activation of Adenosine 5'-Monophosphate-Activated Protein Kinase. *J. Immunol.* **197**, 2891–2899 (2016).
91. S. M. Henson, A. N. Akbar, KLRG1--more than a marker for T cell senescence. *Age* **31**, 285–291 (2009).
92. M. Matsumoto, H. Oshiumi, T. Seya, Antiviral responses induced by the TLR3 pathway. *Rev. Med. Virol.* **21**, 67–77 (2011).
93. J. Dietrich, M. Cella, M. Seiffert, H. J. Bühring, M. Colonna, Cutting edge: signal-regulatory protein beta 1 is a DAP12-associated activating receptor expressed in myeloid cells. *J. Immunol.* **164**, 9–12 (2000).
94. L. L. Lanier, B. C. Corliss, J. Wu, C. Leong, J. H. Phillips, Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature* **391**, 703–707 (1998).
95. L. L. Lanier, B. Corliss, J. Wu, J. H. Phillips, Association of DAP12 with activating CD94/NKG2C NK cell receptors. *Immunity* **8**, 693–701 (1998).
96. P. Delgado, E. Fernández, V. Dave, D. Kappes, B. Alarcón, CD3delta couples T-cell receptor signalling to ERK activation and thymocyte positive selection. *Nature* **406**, 426–430 (2000).
97. O. Barel, Z. Shorer, H. Flusser, R. Ofir, G. Narkis, G. Finer, H. Shalev, A. Nasasra, A. Saada, O. S. Birk, Mitochondrial complex III deficiency associated with a homozygous mutation in UQCRQ. *Am. J. Hum. Genet.* **82**, 1211–1216 (2008).
98. C. Liang, S. Zhang, D. Robinson, M. V. Ploeg, R. Wilson, J. Nah, D. Taylor, S. Beh, R. Lim, L. Sun, D. M. Muoio, D. A. Stroud, L. Ho, Mitochondrial microproteins link metabolic cues to respiratory chain biogenesis. *Cell Rep.* **40**, 111204 (2022).
99. S. Zong, M. Wu, J. Gu, T. Liu, R. Guo, M. Yang, Structure of the intact 14-subunit human cytochrome c oxidase. *Cell Res.* **28**, 1026–1034 (2018).
100. R. Szklarczyk, B. F. J. Wanschers, L. G. Nijtmans, R. J. Rodenburg, J. Zschocke, N. Dikow, M. A. M. van den Brand, M. G. M. Hendriks-Franssen, C. Gilissen, J. A. Veltman, M. Nooteboom, W. J. H. Koopman, P. H. G. M. Willems, J. A. M. Smeitink, M. A. Huynen, L. P. van den Heuvel, A mutation in the FAM36A gene, the human ortholog of COX20, impairs cytochrome c oxidase assembly and is associated with ataxia and muscle hypotonia. *Hum. Mol. Genet.* **22**, 656–667 (2013).
101. S. Hirayama, Y. Yamazaki, A. Kitamura, Y. Oda, D. Morito, K. Okawa, H. Kimura, D. M. Cyr, H. Kubota, K. Nagata, MKKS is a centrosome-shuttling protein degraded by disease-causing mutations via CHIP-mediated ubiquitination. *Mol. Biol. Cell* **19**, 899–911 (2008).
102. C. Regnard, S. Audebert, Desbruyères, P. Denoulet, B. Eddé, Tubulin polyglutamylase: partial purification and enzymatic properties. *Biochemistry* **37**, 8395–8404 (1998).
103. M. Genova, L. Grycova, V. Puttrich, M. M. Magiera, Z. Lansky, C. Janke, M. Braun, Tubulin polyglutamylation differentially regulates microtubule-interacting proteins. *EMBO J.* **42**, e112101 (2023).
104. N. Peng, F. Nakamura, Microtubule-associated proteins and enzymes modifying tubulin. *Cytoskeleton* **80**, 60–76 (2023).
105. H. Doodhi, A. E. Prota, R. Rodríguez-García, H. Xiao, D. W. Custar, K. Bargsten, E. A. Katrukha, M. Hilbert, S. Hua, K. Jiang, I. Grigoriev, C.-P. H. Yang, D. Cox, S. B. Horwitz, L. C. Kapitein, A. Akhmanova, M. O. Steinmetz, Termination of Protofilament Elongation by Eribulin Induces Lattice Defects that Promote Microtubule Catastrophes. *Curr. Biol.* **26**, 1713–1721 (2016).
106. T. Oda, K. Kayukawa, H. Hagiwara, H. T. Yudate, Y. Masuho, Y. Murakami, T. A. Tamura, M. A. Muramatsu, A novel TATA-binding protein-binding protein, ABT1, activates basal transcription and has a yeast homolog that is essential for growth. *Mol. Cell. Biol.* **20**, 1407–1418 (2000).

107. T. Oda, A. Fukuda, H. Hagiwara, Y. Masuho, M.-A. Muramatsu, K. Hisatake, T. Yamashita, ABT1 associated protein (ABTAP), a novel nuclear protein conserved from yeast to mammals, represses transcriptional activation by ABT1. *J. Cell. Biochem.* **93**, 788–806 (2004).
108. K. Dörner, C. Ruggeri, I. Zemp, U. Kutay, Ribosome biogenesis factors-from names to functions. *EMBO J.* **42**, e112699 (2023).
109. S. Singh, A. Vanden Broeck, L. Miller, M. Chaker-Margot, S. Klinge, Nucleolar maturation of the human small subunit processome. *Science* **373**, eabj5338 (2021).
110. H. Zhao, Z. Khan, C. J. Westlake, Ciliogenesis membrane dynamics and organization. *Semin. Cell Dev. Biol.* **133**, 20–31 (2023).
111. N. Kumar, A. K. Singh, The anatomy, movement, and functions of human sperm tail: an evolving mystery. *Biol. Reprod.* **104**, 508–520 (2021).
112. H. Yukitake, M. Furusawa, T. Taira, S. M. M. Iguchi-Ariga, H. Ariga, AAT-1, a novel testis-specific AMY-1-binding protein, forms a quaternary complex with AMY-1, A-kinase anchor protein 84, and a regulatory subunit of cAMP-dependent protein kinase and is phosphorylated by its kinase. *J. Biol. Chem.* **277**, 45480–45492 (2002).
113. G. Martinez, J. Beurois, D. Dacheux, C. Cazin, M. Bidart, Z.-E. Kherraf, D. R. Robinson, V. Satre, G. Le Gac, C. Ka, I. Gourlaouen, Y. Fichou, G. Petre, E. Dulioust, R. Zouari, N. Thierry-Mieg, A. Touré, C. Arnoult, M. Bonhivers, P. Ray, C. Coutton, Biallelic variants in encoding CFAP91, a calmodulin-associated and spoke-associated complex protein, cause severe astheno-teratozoospermia and male infertility. *J. Med. Genet.* **57**, 708–716 (2020).
114. F. Bontems, R. J. Fish, I. Borlat, F. Lembo, S. Chocu, F. Chalmel, J.-P. Borg, C. Pineau, M. Neerman-Arbez, A. Bairoch, L. Lane, C2orf62 and TTC17 are involved in actin organization and ciliogenesis in zebrafish and human. *PLoS One* **9**, e86476 (2014).
115. M. Arafat, A. Harlev, I. Har-Vardi, E. Levitas, T. Priel, M. Gershoni, C. Searby, V. C. Sheffield, E. Lunenfeld, R. Parvari, Mutation in (C2orf62) causes oligoteratoasthenozoospermia by affecting actin dynamics. *J. Med. Genet.*, doi: 10.1136/jmedgenet-2019-106825 (2020).
116. Y. Sha, W. Liu, S. Li, L. V. Osadchuk, Y. Chen, H. Nie, S. Gao, L. Xie, W. Qin, H. Zhou, L. Li, Deficiency in AK9 causes asthenozoospermia and male infertility by destabilising sperm nucleotide homeostasis. *EBioMedicine* **96**, 104798 (2023).
117. B. H. M. Meldal, H. Bye-A-Jee, L. Gajdoš, Z. Hammerová, A. Horácková, F. Melicher, L. Perfetto, D. Pokorný, M. R. Lopez, A. Türková, E. D. Wong, Z. Xie, E. B. Casanova, N. Del-Toro, M. Koch, P. Porras, H. Hermjakob, S. Orchard, Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2019).
118. K. Izumikawa, H. Ishikawa, H. Yoshikawa, S. Fujiyama, A. Watanabe, H. Aburatani, H. Tachikawa, T. Hayano, Y. Miura, T. Isobe, R. J. Simpson, L. Li, J. Min, N. Takahashi, LYAR potentiates rRNA synthesis by recruiting BRD2/4 and the MYST-type acetyltransferase KAT7 to rDNA. *Nucleic Acids Res.* **47**, 10357–10372 (2019).
119. E. Balogh, J. C. Chandler, M. Varga, M. Tahoun, D. K. Menyhárd, G. Schay, T. Goncalves, R. Hamar, R. Légrádi, Á. Szekeres, O. Gribouval, R. Kleta, H. Stanescu, D. Bockenhauer, A. Kerti, H. Williams, V. Kinsler, W.-L. Di, D. Curtis, M. Kolatsi-Joannou, H. Hammid, A. Szőcs, K. Perczel, E. Maka, G. Toldi, F. Sava, C. Arrondel, M. Kardos, A. Fintha, A. Hossain, F. D'Arco, M. Kaliakatsos, J. Koeglmeier, W. Mifsud, M. Moosajee, A. Faro, E. Jávorszky, G. Rudas, M. H. Saied, S. Marzouk, K. Kelen, J. Götze, G. Reusz, T. Tulassay, F. Dragon, G. Mollet, S. Motameny, H. Thiele, G. Dorval, P. Nürnberg, A. Perczel, A. J. Szabó, D. A. Long, K. Tomita, C. Antignac, A. M. Waters, K. Tory, Pseudouridylation defect due to and mutations causes nephrotic syndrome with cataracts, hearing impairment, and enterocolitis. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15137–15147 (2020).
120. T. Vulliamy, R. Beswick, M. Kirwan, A. Marrone, M. Digweed, A. Walne, I. Dokal, Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8073–8078 (2008).
121. C. Wang, U. T. Meier, Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J.* **23**, 1857–1867 (2004).
122. T. H. D. Nguyen, J. Tam, R. A. Wu, B. J. Greber, D. Toso, E. Nogales, K. Collins, Cryo-EM structure of substrate-bound human telomerase holoenzyme. *Nature* **557**, 190–195 (2018).
123. L. Chen, C. M. Roake, A. Freund, P. J. Batista, S. Tian, Y. A. Yin, C. R. Gajera, S. Lin, B. Lee, M. F.

Pech, A. S. Venteicher, R. Das, H. Y. Chang, S. E. Artandi, An Activity Switch in Human Telomerase Based on RNA Conformation and Shaped by TCAB1. *Cell* **174**, 218–230.e13 (2018).

124. C. J. Webb, V. A. Zakian, Telomerase RNA is more than a DNA template. *RNA Biol.* **13**, 683–689 (2016).
125. J. Ge, D. A. Rudnick, J. He, D. L. Crimmins, J. H. Ladenson, M. Bessler, P. J. Mason, Dyskerin ablation in mouse liver inhibits rRNA processing and cell division. *Mol. Cell. Biol.* **30**, 413–422 (2010).
126. J. P. Girard, H. Lehtonen, M. Caizergues-Ferrer, F. Amalric, D. Tollervey, B. Lapeyre, GAR1 is an essential small nucleolar RNP protein required for pre-rRNA processing in yeast. *EMBO J.* **11**, 673–682 (1992).
127. J. S. P. Mawer, J. Massen, C. Reichert, N. Grabenhorst, C. Mylonas, P. Tessarz, Nhp2 is a reader of H2AQ105me and part of a network integrating metabolism with rRNA synthesis. *EMBO Rep.* **22**, e52435 (2021).
128. T. Kinoshita, Biosynthesis and biology of mammalian GPI-anchored proteins. *Open Biol.* **10**, 190290 (2020).
129. K. Kajiwara, R. Watanabe, H. Pichler, K. Ihara, S. Murakami, H. Riezman, K. Funato, Yeast ARV1 is required for efficient delivery of an early GPI intermediate to the first mannosyltransferase during GPI assembly and controls lipid flow from the endoplasmic reticulum. *Mol. Biol. Cell* **19**, 2069–2082 (2008).
130. A. Ikeda, K. Kajiwara, K. Iwamoto, A. Makino, T. Kobayashi, K. Mizuta, K. Funato, Complementation analysis reveals a potential role of human ARV1 in GPI anchor biosynthesis. *Yeast* **33**, 37–42 (2016).
131. H. Okai, R. Ikema, H. Nakamura, M. Kato, M. Araki, A. Mizuno, A. Ikeda, P. Renbaum, R. Segel, K. Funato, Cold-sensitive phenotypes of a yeast null mutant of ARV1 support its role as a GPI flippase. *FEBS Lett.* **594**, 2431–2439 (2020).
132. J. Poli, S. M. Gasser, M. Papamichos-Chronakis, The INO80 remodeller in transcription, replication and repair. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).
133. S. Eustermann, K. Schall, D. Kostrewa, K. Lakomek, M. Strauss, M. Moldt, K.-P. Hopfner, Structural basis for ATP-dependent chromatin remodelling by the INO80 complex. *Nature* **556**, 386–390 (2018).
134. Q. Peng, D. Wan, R. Zhou, H. Luo, J. Wang, L. Ren, Y. Zeng, C. Yu, S. Zhang, X. Huang, Y. Peng, The biological function of metazoan-specific subunit nuclear factor related to kappaB binding protein of INO80 complex. *Int. J. Biol. Macromol.* **203**, 176–183 (2022).
135. E. Cox, W. Hwang, I. Uzoma, J. Hu, C. M. Guzzo, J. Jeong, M. J. Matunis, J. Qian, H. Zhu, S. Blackshaw, Global Analysis of SUMO-Binding Proteins Identifies SUMOylation as a Key Regulator of the INO80 Chromatin Remodeling Complex. *Mol. Cell. Proteomics* **16**, 812–823 (2017).
136. S. Lukauskas, A. Tvardovskiy, N. V. Nguyen, M. Stadler, P. Faull, T. Ravnsborg, B. Özdemir Aygenli, S. Dornauer, H. Flynn, R. G. H. Lindeboom, T. K. Barth, K. Brockers, S. M. Hauck, M. Vermeulen, A. P. Snijders, C. L. Müller, P. A. DiMaggio, O. N. Jensen, R. Schneider, T. Bartke, Decoding chromatin states by proteomic profiling of nucleosome readers. *Nature* **627**, 671–679 (2024).
137. S. M. Reed, J. Hagen, V. P. Muniz, T. R. Rosean, N. Borcherding, S. Sciegienka, J. A. Goeken, P. W. Naumann, W. Zhang, V. S. Tompkins, S. Janz, D. K. Meyerholz, D. E. Quelle, NIAM-deficient mice are predisposed to the development of proliferative lesions including B-cell lymphomas. *PLoS One* **9**, e112126 (2014).
138. D. Gogendeau, M. Lemullois, P. Le Borgne, M. Castelli, A. Aubusson-Fleury, O. Arnaiz, J. Cohen, C. Vesque, S. Schneider-Maunoury, K. Bouhouche, F. Koll, A.-M. Tassin, MKS-NPHP module proteins control ciliary shedding at the transition zone. *PLoS Biol.* **18**, e3000640 (2020).
139. J. Gonçalves, L. Pelletier, The Ciliary Transition Zone: Finding the Pieces and Assembling the Gate. *Mol. Cells* **40**, 243–253 (2017).
140. C. Li, V. L. Jensen, K. Park, J. Kennedy, F. R. Garcia-Gonzalo, M. Romani, R. De Mori, A.-L. Bruel, D. Gaillard, B. Doray, E. Lopez, J.-B. Rivière, L. Faivre, C. Thauvin-Robinet, J. F. Reiter, O. E. Blacque, E. M. Valente, M. R. Leroux, MKS5 and CEP290 Dependent Assembly Pathway of the Ciliary Transition Zone. *PLoS Biol.* **14**, e1002416 (2016).